Do Google Search Trends Allow us to Better Understand Residential Adoption of Solar?

Student Team: *Qiuying Lai, Ghida Ismail, Ryan Williams*
Internal Technical Lead: *Dr. Ashok Sekar*
Advisor and Project Director: *Dr. Varun Rai*

Policy Research Project (PRP)
LBJ School of Public Affairs
The University of Texas at Austin
May 2018

Abstract

Researchers have mined data on people's online activities to "nowcast" economic or market trends for the past 10 years. In this project, the team tested whether Google search trends can nowcast residential PV adoption. Specifically, the team conducted an empirical analysis to assess the explanatory power of Google search trends for two solar PV markets in U.S.—California and Connecticut—over the years 2004 to 2016.

Of the large variety of search terms explored (~100 different terms were tested), "solarcity" performed the best in predicting subsequent adoption (other terms that also performed adequately include "solar cost", "Sunrun", "solar panel", and "solar tax credit"). More specifically, searches for "solarcity" explained much of the variability in month-to-month statewide residential solar adoption in California, though less so in Connecticut. In addition, the research identified a lag between search trends and actual adoption (4-7 months), likely reflecting the natural lag between information search and actual adoption. However, the term "solarcity" became insignificant in predicting adoption when included in a regression model that contained other relevant explanatory factors of adoption such as incentives and socio-demographic variables. This suggests that one might use search trends as a replacement for these other variables, but that search trends alone may not independently add much additional ability to nowcast solar adoption. This finding was confirmed by the high correlation coefficient (>|0.5|) of the search term "solarcity" with these other variables.

Note that SolarCity (the installer) has merged with Tesla, so the search term "solarcity" may not be a good predictor of PV adoption (even in California) going forward. Nonetheless, the findings suggest that Google search trends have the potential to help nowcast residential solar PV adoption, perhaps especially in cases where other relevant correlates are not readily available. Further analysis, including more states and adding 2017 adoption data, is needed to more fully understand the contours of the potential. Furthermore, data permitting, a more granular analysis (e.g., at the county level) may shed additional insights.

Table of Contents

INTRODUCTION

Are the preferences and behaviors of consumers influenced or revealed by online activity? Internet search and social media activity, particularly the study of internet search terms and Twitter activity data, has been linked to real-time health-trends and economic activity. Notable examples include internet search trends that have been found to be accurate predictors and detectors of influenza epidemics[1], using search queries from Yahoo and Google as well as Twitter trends. Search term queries have also been used to "nowcast" – or predict the present – macro-economic indicators. As indicators such as GDP growth or unemployment rates often are released with a considerable lag time as data is collected, prepared, and analyzed, economists can use search term trends to create more robust models to estimate economic indicators[2]. Search term volume can also be used to forecast current sales of consumer purchases such as box office revenue or video game sales[3].

Furthermore, as the usage of the internet and social media is growing in today's society, Internet searches and social media activity could be used to take a snapshot of consumer sentiment, opinion and awareness of solar. Previous work has found that internet searches are very good indicators of consumer sentiment, when compared with survey-based consumer sentiment indices[4]. Internet search data have also been used to suggest

[1] Polgreen, Philip M., Yiling Chen, David M. Pennock, and Forrest D. Nelson. 2008. "Using Internet Searches for Influenza Surveillance."Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America47 (11): 1443–48. doi:10.1086/593098.

[2] Vosen, Simeon, and Torsten Schmidt. 2011. "Forecasting Private Consumption: Survey-Based Indicators vs. Google Trends."Journal of Forecasting30 (6): 565–78. doi:10.1002/for.1213.

[3] Goel, Sharad, Jake M. Hofman, Sébastien Lahaie, David M. Pennock, and Duncan J. Watts. 2010. "Predicting Consumer Behavior with Web Search." Proceedings of the National Academy of Sciences107 (41): 17486–90. doi:10.1073/pnas.1005962107.

[4] Della Penna, Nicolás, and Haifang Huang. 2010. "Constructing Consumer Sentiment Index for U.S. Using Google Searches." Working Paper 2009–26. University of Alberta, Department of Economics.http://econpapers.repec.org/paper/risalbaec/2009_5f026.htm.

trends in public interest in specific areas, such as the environment[5]. Therefore, the data generated from Internet activities has the potential to provide fundamental insights on people's perception and behavior towards solar technologies which can inform the relevant stakeholders.

Building on the literature cited above, this paper addresses the question of whether internet activity plays a role in the prediction of the diffusion of PV at different geographic and time scales in the United States. Prediction of PV Diffusion has become an essential issue as it enables solar industry firms to identify efficiency loss or plan for systems distribution. The ability to forecast PV diffusion is also vital for utilities and other grid planners who require knowledge of how the grid is developing. Ideally, this kind of prediction would be constantly updated with the most recent streams of relevant data. Unfortunately, this data is not updated in anything close to the ideal- realtime- that solar firms and grid planners desire. This is where our research makes a contribution.

The most robust database for national, state, and local PV adoption time series is released with a delay of 6 to 12 months. If internet activities are proven to have a significant predictive impact on the diffusion of PV, it will enable the provision of more prompt and timely information on PV adoption. Existing models aiming at forecasting PV installations are usually very complex or require a lot of data that is sometimes hard to collect or outdated. Therefore, finding simpler methods to nowcast PV adoption can be very beneficial for the solar industry.

Our primary research objective in this study is to determine how, if at all, google search activities correlate with the residential adoption of PV panels and whether google

[5] Mccallum, Malcolm L., and Gwendolyn W. Bury. 2013. "Google Search Patterns Suggest Declining Interest in the Environment." Biodiversity and Conservation22 (6–7): 1355–67. doi:10.1007/s10531-013-0476-6.

searches can be effective proxies for other variables that impact consumers' willingness to adopt Solar Panels, such as peer effects, the entry of installers to the market or other types of unobserved variables. Subsequently the study aims at determining the role of google searches in predicting PV Diffusion and provide future researchers' grounds to understand what role the internet plays in predicting the diffusion of Solar Panels.

To better grasp the function of google searches in different markets in the United States, we focus our analysis on two states, California, with a large PV market and Connecticut with a small market.

We first compile descriptive statistics of PV Diffusion and on Google searches activities. Then we identify google search terms highly correlated to PV diffusion. Subsequently we use the Tracking the Sun datasets for the data on PV installations data, google searches retrieved from Google Trends[6] and socio-demographic data retrieved from the American Community Survey[7] to run empirical models evaluating the predictive power of the google searches.

DATA AND DATA PROCESSING

Solar PV installation data

We use the Tracking the Sun dataset for the solar PV installation data retrieved from the Open PV Project at the National Renewable Energy Laboratory (NREL). They record the grid-connected PV systems installed across the different states in the US from 1998 to the end of 2016. However, for this analysis we use only residential data ranging in size from 1KW to 10 KW and installed between 2004 and 2016. This data includes 573,434 residential systems in California and 15,581 in Connecticut.

[6] trends.google.com. 2018

[7] US Census Bureau. 2018. https://www.census.gov/programs-surveys/acs/

In the dataset, each observation represents one installation. Considering that we can only access google search volume at the monthly level, our analysis will be conducted at the monthly level. Therefore, we create a PV Diffusion variable recording the number of installations per month for each of California and Connecticut. Furthermore, we transform the installation price, rebate amount, Sales Tax Cost, Performance Based Incentives and Feed in Tariff Annual Payment to real 2016 prices using the World Bank's Consumer Index Price[8]. We then normalize them by their respective system size and compute their average per month.

Google trends data

We use a list of google search terms that reveal consumers' interest in adopting Solar/ PV Panels. This is explained in greater details in the next section.

We collect data on the search volume of these terms by using Google trends data and a Python Script adapted from an unofficial API called "pytrends". Given a specific Region and time range, Google Trends enable us to access normalized percentage of searches. The numbers in the Google Trends data indicate the search interest as a percentage of the highest search point in the time period and region chosen. In other words, a value of 100 means the peak volume of search happened at this point in the specified region and time range, while a value of 0 reveals that the search term represents less than 1% of the peak search volume.

We extract the data in one round from 2004 through 2016 for each of California and Connecticut, instead of extracting it on a per year basis to make sure that the Google Trends data is consistently normalized over the whole period of study. In other words, a

[8] Data.worldbank.org, 2018

value of a 100 in the data we extracted represents the maximum volume of searches over the whole period from 2004 through 2016. It is important to acknowledge here that working with normalized data might limit the accuracy of the analysis.

Baseline Data

We collect average yearly residential electricity prices in Cents/Kilowatt-hour for California and Connecticut from 2005 to 2016 from the US Energy Information Administration (EIA) using the form EIA-861.

On the other hand, we amass the percentage of population 25 years and over with a Bachelor's degree and the percentage of population in different age demographic using demographic data from the American Community Survey (ACS). The ACS is a statistical survey conducted by the US Census Bureau that samples a small percentage of the US population every year aiming at providing communities with demographic, housing, social, and economic data. The collected variables for each year from 2005 through 2016 are as follows:

- *Total; Estimate; Population 25 years and over - Bachelor's degree* from the S1501 form.
- *Percent; SEX AND AGE* - 20 to 24 years, 25-35 years, 35-44 years, 45-54 years, 55-59 years, 60-64 years, 65 -74 years, 75- 84 years and 85 years and older from the DP05 form. We then collapse these variables into 10 years age groups instead of keeping them as 5 years groups.

We also use data on median income in 2016 dollars from the US Census Bureau, specifically Table H-8- Median Household Income by State: 1984 to 2016 and data on homeownership using Table 3- Homeownership Rates by State: 2005-present from the

Quarterly Vacancy and Homeownership Rates by State and MSA. We use the homeownership rate in the first quarter of each years from 2005 through 2016.

DESCRIPTIVE ANALYSIS

We use the PV installations data to build an understanding of underlying trends in residential PV diffusion. We plot the yearly variation of the count of residential PV systems in California and Connecticut, which reveal a notable upward trend in both states as shown in Figure 1. We note that although they have similar trends, the PV installations in California is about 150 times more than that in Connecticut. They both peak at the end of 2015 and then starts decreasing.

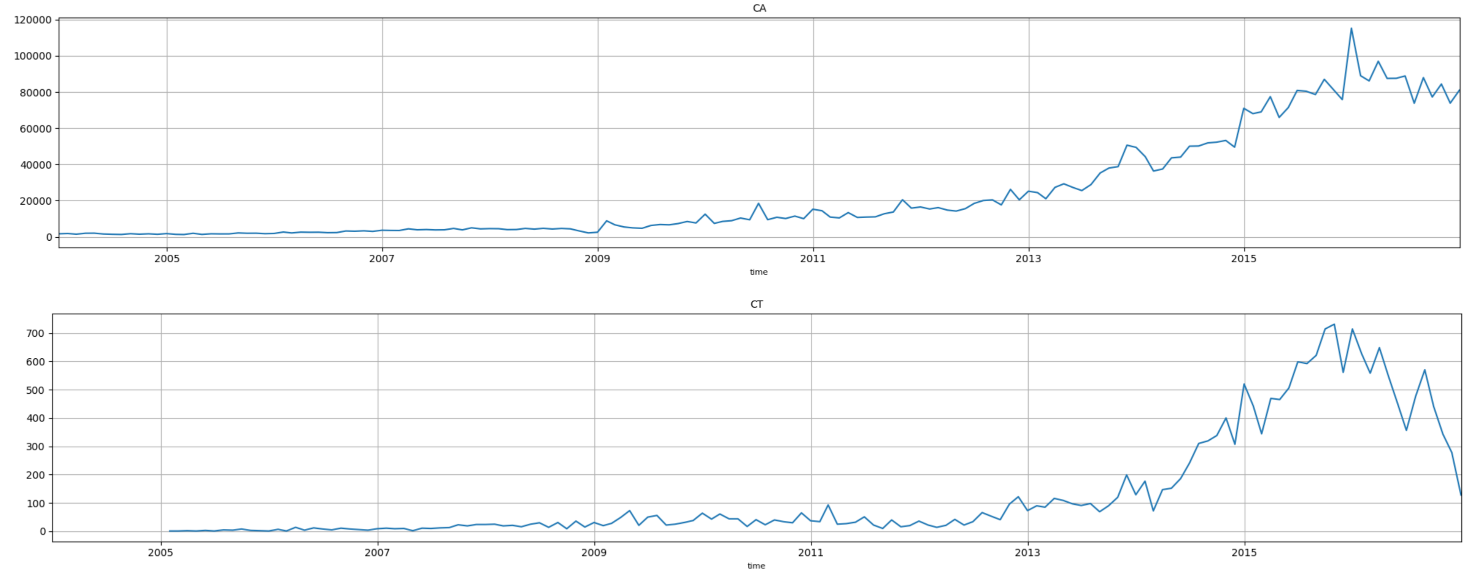

Figure 1: Time series of residential PV installation count in CA and CT

We characterize the leading companies in the California residential solar market by the average cost per watt of the systems they install, and the total number of those systems as shown in Figure 2. This technique can be used to identify market leaders or innovative companies, whose searches of names might serve as variables in a predictive model.

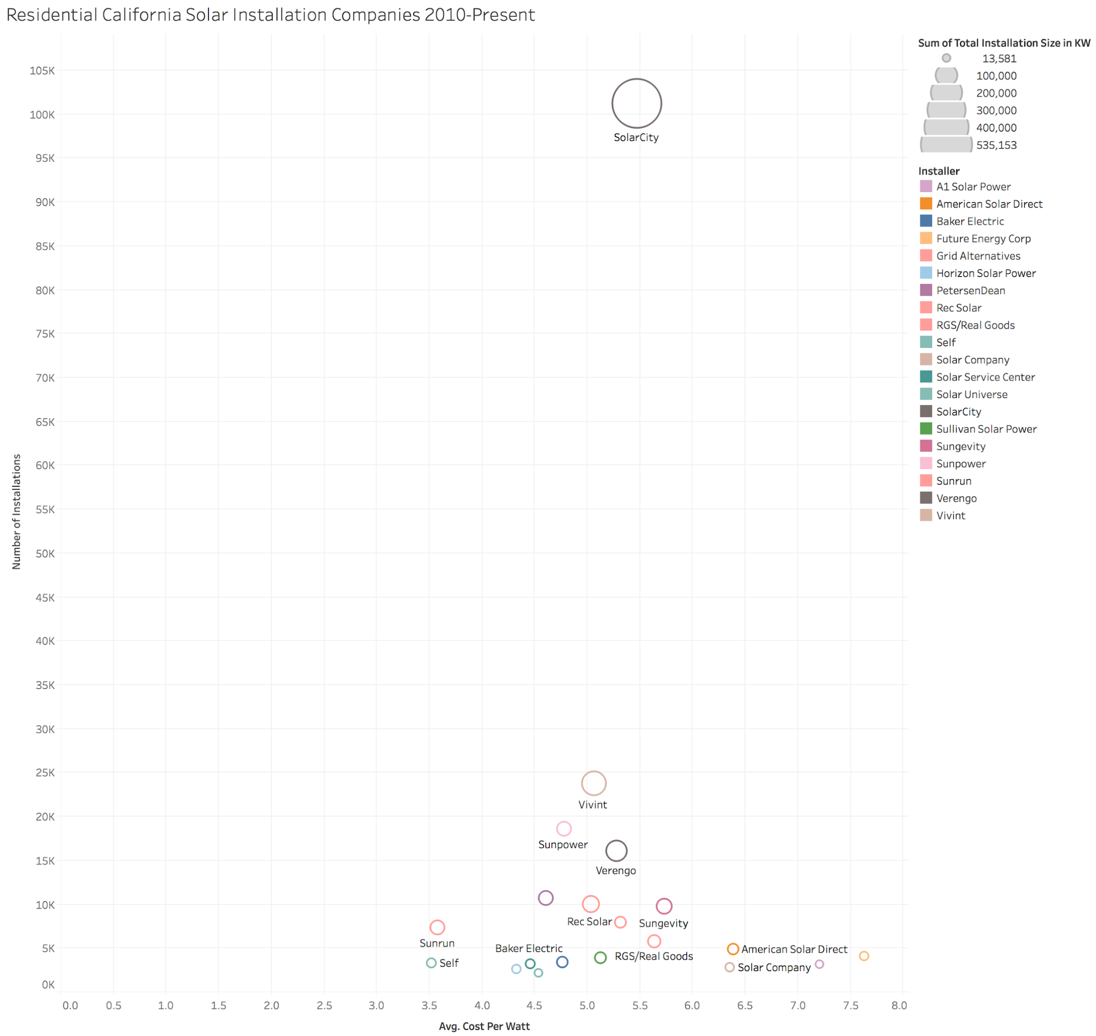

Figure 2: Residential Solar Installation Companies in California

Moreover, we paired the Google Trends data with demographic indicators, energy pricing data, and net metering scores, with the hope that we might acquire some sense of the underlying variation in google trends data between states and how that variation might correlate with other salient data points, especially data that has been used before in PV diffusion models. This descriptive analysis not only reveals which variables the google trends data act as proxies for, they also reveal how google trends data, demography, and residential PV installations vary across time.

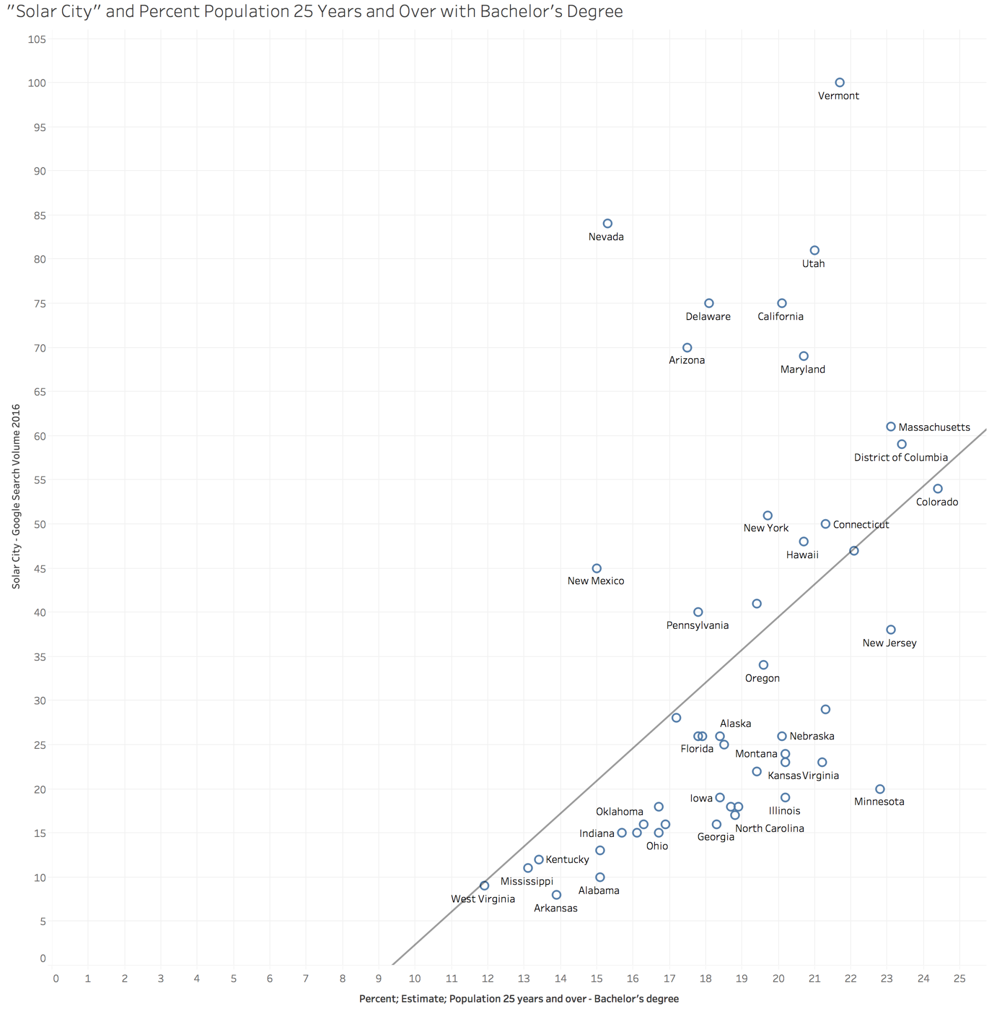

Figure 3: "SolarCity" and Percent Population 25 Years and Over with Bachelor's Degree (2016)

One of the most striking results of the descriptive analysis was the correlation between the search terms for which there was enough data in 2016 and the level of education as operationalized by census data on the percent of population 25 years and older with a Bachelor's degree. The trend line featured in the figure above has an R-squared of .21 at a high degree of significance. It is also clear, however, from the discontinuity between the topmost cluster of states that there are other powerful drivers of search behavior. The correlation is even stronger with searches for "solar cost". The R-squared in this case is .27. This suggests that, for at least the year of 2016, “solar cost” might perform slightly better than “solarcity” as a proxy for educational attainment.

We examined a great number of combinations of search terms and demographic variables. It is interesting to note that many semantically related terms have very different relationships to certain variables. For instance, there is a very clear positive correlation between the average price of electricity per state and the search term “solar installation” and basically no relationship between electricity price and the search term “solar panel cost”. Statistically significant results for the most successful search term, “solarcity”, are summarized in Table 1.

The results of this analysis were useful to adjust our intuitions about certain propositions. For instance, by clearly laying out the results visually, we can have a better understanding of the relationships between our search terms and other demographic variables traditionally associated with PV diffusion. This kind of analysis can be used to explore the correlation between search data and demographic variables in one state across time. More of this kind of analysis could provide future researchers with a better toolkit for generating and evaluating search term derived proxies of PV diffusion determinants.

Table 1: Correlation Between “SolarCity” and Various Indicators

Indicator	R-Squared
Median Household Income	.28
Average Price of Electricity	.15
Percent of Population 25 Years and Older with Bachelor’s Degree or Higher	.21
Percent of Labor Force Currently Employed	.09

We also plot the variation of searches for the term "Solarcity" by month for the year 2016, for California, Connecticut and the US, to identify any interesting seasonal trends, as shown in Figure 4. We note that the search trends in the US and California are very similar, which might indicate that California is driving the searches for Solarcity in the US. The highest peak of searches in both California and Connecticut is during the summer. In California the searches peak in June, while in Connecticut searches peak in July. This might imply that adopters' interest in solar panel peaks in the summer. This analysis also shows how google searches can reflect seasonality in adopter's sentiments and interest in solar panel.

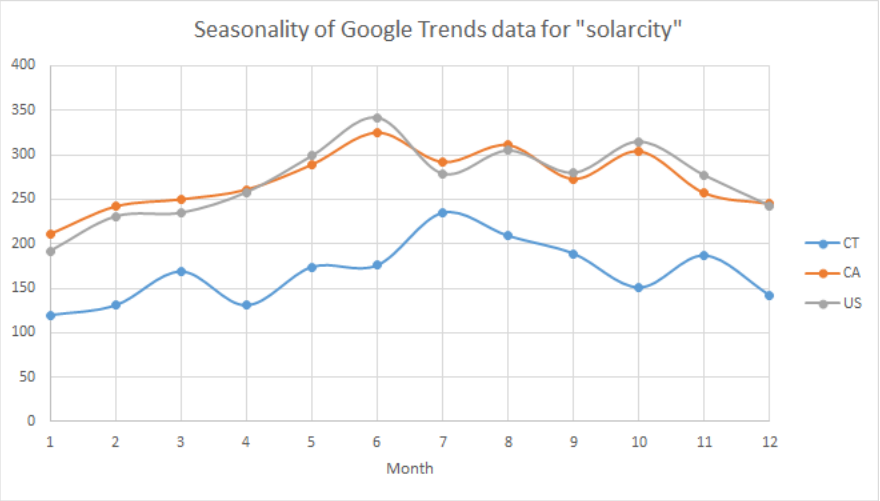

Figure 4: Seasonality of Searches for Solarcity

IDENTIFICATION OF GOOGLE SEARCH TERM PREDICTOR

We choose search terms that might indicate people's interest in adopting PV. This initial collection of search terms was generated by imagining what residential PV adopters might search for in the months before installation. This includes searches for cost of solar panels, installations as well as names of major installers. We also use Google Correlate[9] to generate a list of secondary solar related terms, which means we initially set “solar” as an input and get 100 correlated terms for the period of 2004-2016. Then those 100 terms are used as input and their correlated terms are returned.

The final full list of terms used is shown in Table 2.

Table 2: Google Search terms

Key Terms			
average PV cost	solar world usa	solar jobs	PV installation
avg PV cost	sun power	solar panel	PV price
buy solar panels	solar panel cost	solar panel companies	PV rebate
comparison PV	solar panel kits	solar panel installation	PV roof
comparison solar	solar panel calculator	Solar Price	PV tax credit
first time pv	solar panel tax credit	solar rebate	Solar calculator
first time solar	solar panel efficiency	solar rebates	solar companies
how to install solar panels	solar panel prices	solar roof	Solar contractors
how to install solar	solar panel system	solar sales	Solar Cost
installing solar panels	solar panel companies	Solar tax credit	solar finance
solar incentives	Net metering	SolarCity	sunrun
local solar	Solar Cheap	Sunroof	vivint
New home PV	Solar expensive	first solar	trinity solar
New home solar	Solar panel cheap	solar universe	
PV calculator	solar panel expensive	Solar installation	
PV cost	install solar panel	Solar Installer	

We run pairwise correlations between monthly PV adoption in both California and Connecticut and the above chosen search terms. We assume that google search happens in the exploratory phase defined as a pre-decision-making phase. The decision to install solar

[9] Google Correlate.2018. https://www.google.com/trends/correlate

panels does not lead to an instantaneous installation, due to the time needed to complete the necessary paperwork and perform the installation. Therefore, there's a time between the google searching and the actual PV installation. In order to account for that, we run correlations between monthly PV adoption and the Google search terms at lags ranging from 1 to 9 months.

The terms with the highest correlations with their respective lags are shown for California and Connecticut in Table 3 and 4 respectively. For Connecticut most terms in the list in Table 2 did not have search volumes. It is interesting to note that for both California and Connecticut the terms with the highest correlation with monthly PV adoption are names of major installers. For California, "Sunrun" and "Solarcity" are leading the list with correlations coefficients of 0.9 and 0.88 respectively. For Connecticut, "Solarcity" is also leading with a correlation coefficient of 0.79, while all other terms have very low correlation coefficients at around 0.3.

Table 3: Correlation Coefficients California

State	Term	Time Between Search and Adoption (Months)	Correlations Coefficients
California	Sunrun	4	0.9
	SolarCity	7	0.88
	Solar Cost	7	0.75
	Solar Panel	5	0.75
	Solar Tax Credit	3	0.75

Table 4: Correlation Coefficients Connecticut

State	Term	Time Between Search and Adoption (Months)	Correlations Coefficients
Connecticut	SolarCity	5	0.79
	Solar Companies	6	0.27
	Solar Cost	5	0.3

Based on these results we plot for both California and Connecticut the monthly variation of searches for “Solarcity” and the monthly PV adoption 7 months later for California and 5 months later for Connecticut. We also plot for California, the monthly variation of searches for “Sunrun” and the monthly PV adoption 4 months later and the monthly variation of searches of “Solar Panel” and the monthly PV adoption 5 months later, as shown in Figures 4, 5, 6 and 7. The numbers on the x-axis of the plots represent the number of months with the first month being January 2004, the 12th month is December 2004 while the 144th month is January 2016.

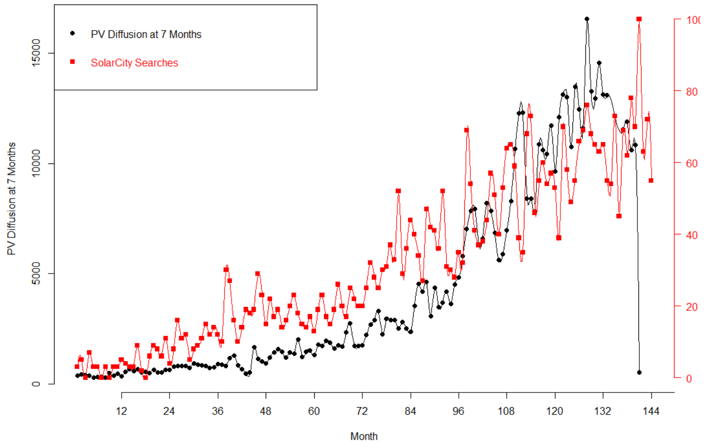

Figure 5: PV Diffusion at 7 months and Searches of “Solarcity” in California

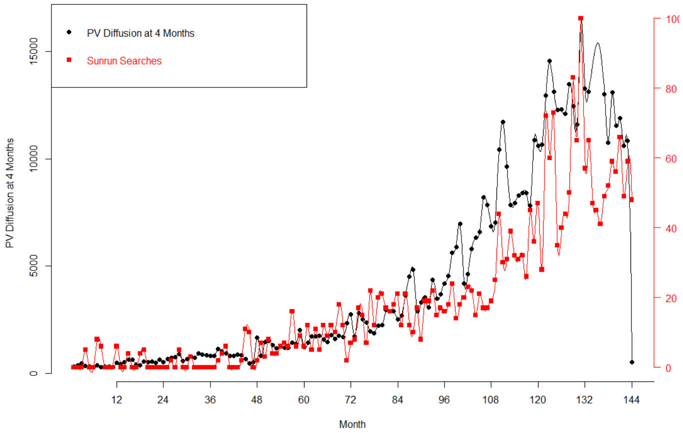

Figure 6: PV Diffusion at 4 months and Searches of “Sunrun” in California

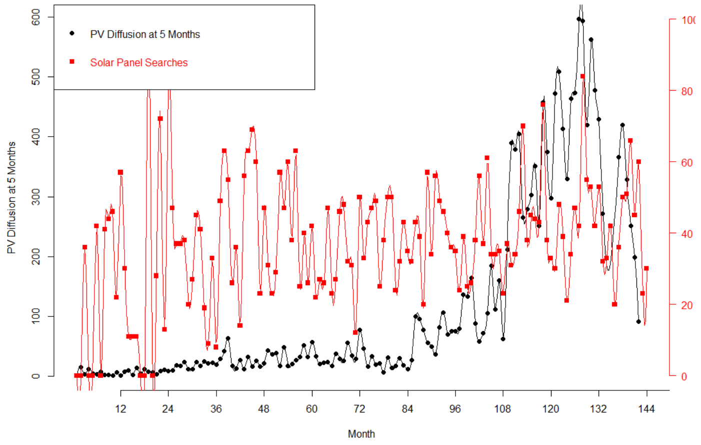

Figure 8: PV Diffusion at 5 months and Searches of "Solar Panel" in California

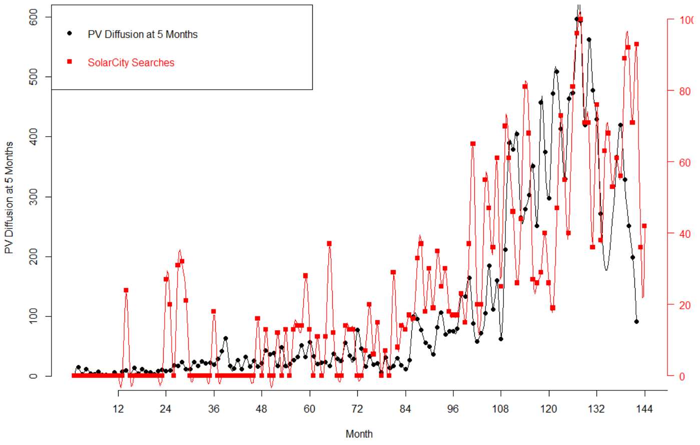

Figure 7: PV Diffusion at 5 months and Searches of "SolarCity" in Connecticut

As suggested by the high coefficients of correlations, these figures reveal that for several months the variation of PV adoption at later time is mirrored by the searches of these terms. The plots also show that the high correlations are not necessarily explained by a general time trend wherein both PV adoption and searches are increasing. In fact, it appears that in several months searches are reflecting both increases and decreases in PV adoption at the monthly level.

In the next section we investigate the predictive power of these three search terms -" Solarcity", "Sunrun", "solar panel" - by running several regressions. We use for California the terms, "Solarcity", "Sunrun" and "solar panel", and for Connecticut we use "Solarcity". We chose to test these terms as they have correlations coefficients of 0.75 and larger.

PREDICTIVE MODELS

We use regressions to evaluate the performance of search-based models relative to baseline models based on publicly available data in California and Connecticut.

Furthermore, we assess whether combining both google searches and baseline predictors improves the prediction of PV diffusion.

Search Based Model

REGRESSIONS

We run two search-based predictions models. We run for California models with “Solarcity”, “Sunrun” and “Solar Panel” at their respective lags, and for Connecticut we run models with “Solarcity”.

The first model pools all years together and is based on a linear model of the form:

$$\text{Log (PV Diffusion}_{\text{t=dmonths}}) = B_0 + B_1\text{searchterm}_i + e_i$$

The response variable PV Diffusion$_{\text{t=dmonths}}$represents the monthly count of installed residential PV systems d months after the google search occured. The d is the number in months identified in the previous section specific for each term. For example, for the term “Solarcity” in California it is equal to 7 months. Furthermore, for the search term “Solarcity” in California we start from the year 2006 and in Connecticut we start at 2012, as these are the years SolarCity opened in these states. Similarly, for “Sunrun” we start from the year 2007, as this is the year of its establishment. It is worth noting here that the results do not vary greatly when including all years from 2004 to 2016.

We log transform the variable PV Diffusion accounting for the fact that it is highly skewed and reducing heteroskedasticity in the model.

The search term is the normalized monthly search volume provided by Google Trends.

The second one is a year fixed effect which allows us to control for heterogeneity and differences across years. Each year might be different from the other as it might include a different political environment, different policies and different incentives and programs to adopt PV.

The equation becomes:

$$\mathrm{Log}(\mathrm{PV\ Diffusion}_{t=dmonths}) = B_0 + B_1 \mathrm{searchterm}_i + B_2 \mathrm{Year} + e_i$$

The vector Year is the year factor variable vector including dummy variables for years to control for potential unobservable yearly changes.

We choose to run two models considering that in many cases, there might not be yearly PV installations data to carry out a year fixed effects model for the prediction. Accounting for the fact that PV installations data are released with a delay of 6 to 12 months, if researches, for example, wish to predict installations in 2018, the installations of the previous year might not be available for a year fixed effects model. It is therefore essential to also evaluate the performance of a pooled model.

We run both models for both California and Connecticut with the identified search terms. We then evaluate the performance of each search-based model by using K fold cross validation. K fold cross validation is a technique that assesses how the results of a model will generalize to an independent data set. It evaluates the predictive power of a model by partitioning the original sample into k roughly equal size subsamples. Then k-1 subsamples are used as training data while one subsample is retained for validation and testing. The process is repeated k times with each subsample as validation data, and then the k measures of fit generated are averaged to produce one single estimate. In our analysis we report the

R2 and Residuals Mean Square Errors (RMSE) as the measure of fit. We choose k to be equal to 10.

RESULTS

In the year pooled models, for California, the terms “Solarcity”, “Sunrun” and “Solar Panel” are all statistically significant at a 99% level of confidence. Similarly for Connecticut, “Solarcity” is statistically significant at a 99% level of confidence.

In the year fixed effects models, for California, only the term “solarcity” is statistically significant at a 99% level of confidence, while Sunrun and Solar Panel are no longer significant. It therefore seems that their significance in the previous model was proxying for a time trend, and when the time trend was controlled for, their significance was lost. For Connecticut, “solarcity” also loses significance when the differences across years is accounted for.

The results of the cross validation of the models for both states are shown in Table 5. Model 1 in the table is the year pooled model, while Model 2 is the year fixed effects.

Table 5: Search Based Models

Search Based Models	**California**				**Connecticut**			
	Model 1		**Model 2**		**Model 1**		**Model 2**	
	RMSE	R2	RMSE	R2	RMSE	R2	RMSE	R2
SolarCity	0.45	85%	0.18	97%	0.69	45%	0.49	89%
Sunrun	0.65	72%	0.2	96%				
Solar Panel	0.89	40%	0.2	96%				

For California, when the years are pooled, the term “solarcity” outperformed the other terms with the lowest RMSE at 0.45, and “sunrun” outperformed the term “solar panel”. Furthermore, the test R2 of 85% for “Solarcity”, 72% for “Sunrun” and 40% for “Solar Panel” mean that the years pooled model with “Solarcity” is able to explain 85 % of the variation of PV Diffusion 7 months later in an independent dataset, the one with “Sunrun” is able to explain 72% of the variation of PV Diffusion 4 months later and the one with “Solar Panel” is able to explain 40 % of the variation of PV Diffusion 5 months later. Therefore, the models based on searches of the names of major installers namely SolarCity and Sunrun perform well in explaining variation in PV adoption when the years are pooled, however the performance of the model based on searches of “solar panel” is much weaker.

In the case of Connecticut, when the years are pooled, the model based on searches of “solarcity” was weaker than all three models in California, with a lower test RMSE at 0.69. Furthermore, the test R2 of 45% signifies that the model is able to explain 45 % of the variation of PV Diffusion 5 months later in an independent dataset. Thus, the search based model has a weak performance in Connecticut. This could be explained by the fact that Connecticut has a small PV market, not enough for search terms of installers to be predictive of adoption.

For the years fixed effects, it is interesting to note that all models performed well with very high R2, and with test RMSE lower than the pooled models. This could be due to the fact that, considering the upward trends of Solar Adoption over the years in both California and Connecticut, yearly factors alone are very powerful in explaining variations in PV adoption. In fact, for the years fixed effect models of “sunrun” and “solar panel” in California, and “Solarcity” in Connecticut, the search term is not statistically significant, which means that the strong performance of the fixed effects model can be attributed to the yearly factors. However, for California, the search term "Solarcity" is still statistically significant even when controlling for yearly factors, which might suggest that the searches for "solarcity" might capture variation in PV adoption not explained by changes in yearly factors.

These results might suggest that searches for major installers such as solarcity and sunrun perform well in predicting PV adoption in large PV market such as California. It also seems that these search terms are proxying for yearly factors.

Baseline Model

Regressions

To put these results in the proper perspective we next compare the search-based predictions with simple models built on publicly available information based on predictors identified in the literature on prediction of PV Diffusion.

In this case we run two models. The first one pools all years together from 2005 to 2016 and is based on a linear model of the form:

$$\text{Log}(\text{PV Diffusion}_{t=dmonths}) = B_0 + B_1\text{InstalledBase} + B_2\text{Socio-Demographic} + B_3\text{Price} + B_4\text{Season} + e_i$$

The Installed Base vector includes PV Diffusion at the time the adopter is searching for "solarcity". Several studies show that previous nearby adoptions affect current PV system adoption[10]. They in fact demonstrate that one additional previous installation in a zip code increases the probability of a new adoption in the same zip code by 0.78% in California.

[10] Among others, see: (A) Rai, V., & Robinson, S. A. (2013). Effective information channels for reducing costs of environmentally-friendly technologies: evidence from residential PV markets. *Environmental Research Letters*, *8*(1), 014044. (B) Rai, V., Reeves, D. C., & Margolis, R. (2016). Overcoming barriers and uncertainties in the adoption of residential solar PV. *Renewable Energy*, *89*, 498-505.

The Price vector includes monthly average installation price, Sales Tax Amount, Performance Based Incentives payment and the Feed in Tariff to account for their role in promoting PV adoption and the financial drivers of PV adoption. All these variables are normalized by size as described in the Data section.

The Socio-Demographic Vector includes the percentage of the population above 25 years old with a Bachelor degree to control for the education and level of awareness in the population and the median income in 2016 dollars to capture overall economic conditions. Also we include the percentage of population in different age demographic, specifically: 20 to 34 years old, 35 to 44 years old, 45 to 64 years old and older than 65 years. In fact, Drury et al. demonstrate that age, income and education are the primary drivers for predicting PV adoption in California[11]. Additionally, we include the yearly statewide percent of homeownership to account for the houses available for installation of PV. We also control for the statewide annual average residential electricity price at the time the adopter is searching for the google terms to control for changes in electricity prices which may affect the interest in PV systems.

The Season vector includes year-quarter dummy variables. It includes a Winter dummy from December to February, a Spring dummy variable from March to May, a Summer dummy from June to August and a Fall dummy from September to November. We include this vector to control for changing trends with season in the PV system market.

The second one is a year fixed effect, and in this model, considering that the year fixed effect accounts for differences across years, we remove the predictors specific for each year.

$$\text{Log}(\text{PV Diffusion}_{t=dmonths}) = B_0 + B_1\text{InstalledBase} + B_2\text{Year} + B_3\text{Price} + B_4\text{Season} + e_i$$

[11] Drury, Easan et al. 2012. The transformation of southern California's residential photovoltaics market through third-party ownership. Energy Policy, Volume 42: 681–690.

The d represents the time in months between decision making and adoption of PV. We try different times of 1 to 9 months and evaluate the performance of these models using a 10 fold cross validation. We chose a time d=6 months as it minimizes the test RMSE of both models for both States.

RESULTS

The results of the 10 fold cross validation of those models at time d=6 months are shown in Table 6.

Table 6: Baseline Models

Baseline Models	**Model 1**		**Model 2**	
	RMSE	R2	RMSE	R2
California	0.28	92%	0.3	92%
Connecticut	0.43	89%	0.46	90%

In the case of California, the baseline pooled model outperformed the search based models for all three terms- “solarcity”, “sunrun” and “solar panel”. These results mean that although searches for “Solarcity” and “sunrun” are predictive, alternative information sources based on publicly available data still perform better. Nonetheless search based models can be useful when baseline data is not available, outdated or hard to collect.

When year fixed effects are considered the search-based models outperform the baseline models. However, this robust performance is mostly due to the yearly factors that are effectively able to explain the variation in PV adoption.

In the case of Connecticut, when the years are pooled the baseline models, with a test RMSE of 0.43 also outperform the search-based models with “solarcity”. As for the year fixed effects, the search based, and the baseline models performed equally well with a test RMSE of around 0.46.

Combined Models

REGRESSIONS

Next, we consider models combining both baseline data and searches of “solarcity”.

We run a pooled year model as follows:

$$\text{Log}(\text{PV Diffusion}_{t=6months}) = B_0 + B_1\text{InstalledBase} + B_2\text{Socio-Demographic} + B_3\text{Price} + B_4\text{Season} + B_5\text{Searchterm} + e_i$$

We also run a year fixed effect model in which we remove the yearly indicators as follows:

$$\text{Log}(\text{PV Diffusion}_{t=6months}) = B_0 + B_1\text{InstalledBase}_i + B_2\text{Year} + B_3\text{Incentives} + B_4\text{Season} + B_5\text{Searchterm} + e_i$$

RESULTS

We note that in the combined models, “Solarcity” and “sunrun” are statistically significant in the years pooled models for California. This might suggest that these search terms are explaining variation in PV adoption that the price and socio-demographic variables have not captured. However, they lose significance in the year fixed effects models. The term “solar panel” is not statistically significant in any of the combined models.

In Connecticut, the term “Solarcity” is not statistically significant in both models.

The results of the cross-validation are shown in Table 7.

Table 7: Combined Models

Search Based Models	California				Connecticut			
	Model 1		Model 2		Model 1		Model 2	
	RMSE	R2	RMSE	R2	RMSE	R2	RMSE	R2
SolarCity	0.27	93%	0.29	92%	0.44	90%	0.47	90%
Sunrun	0.27	93%	0.29	92%				
Solar Panel	0.28	93%	0.3	92%				

We note that the performance of the combined models in both states is identical to the performance of the baseline models. This implies that the search terms and the baseline indicators might be explaining common attribute of PV adoption. These search terms might be instrumenting for these baseline characteristics. We test this hypothesis for the term "solarcity" in the next section.

"SOLARCITY' AS INSTRUMENT

In order to test the hypothesis that "solarcity" is instrumenting for other baseline attributes, we run for California pairwise correlations between "solarcity" and the baseline variables: Median Income, Percent Homeownership, percent population for the different age demographic, percentage of population over 25 years old with a Bachelor's Degree, PV Installed base, PV Price, Electricity Price, Feed in Tariff, Performance Based Incentives and Sales Tax. Figure 8 shows the results of the correlations. It appears that the searches of "SolarCity" is highly correlated to most variables with a Pearson correlation of above 0.75. The variables that appear to be weakly correlated to the searches of "solarcity", with a correlation coefficient of less than 0.5, are the electricity price, median income, the PV price and performance-based incentives. This suggests that "solarcity" might be instrumenting for a combination of the installed PV systems, of socio-demographic factors as measured by percent of homeownership, population age and percent of population above 25 years old with a Bachelor's degree and incentives as measured by Sales Tax cost.

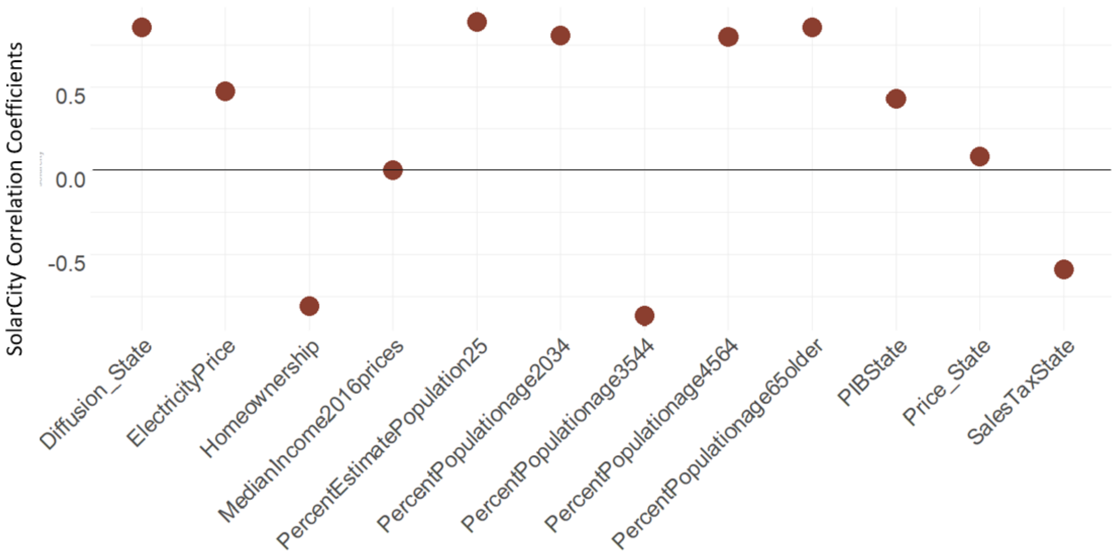

Figure 9: Pairwise Correlations

We further test the performance of "solarcity" as an instrument by running a linear regression of the form:

$$\text{SolarCity}= B_0 + B_1\text{InstalledBase}_i + B_2\text{Socio-Demo} + B_3\text{Incentives} + B_4\text{Season} + e_i$$

With the response variable measuring the normalized searches of SolarCity in California. The Installed base vector includes the installed PV base at the time of the search, the Socio Demographic factor includes fraction of population in different age demographics, percent of homeownership and percent of population above 25 years old with a bachelor's degree. The Incentives vector includes the Sale tax cost. We limited the predictors to the variables that were correlated to the searches of "Solarcity" with a Pearson correlation coefficient of higher than 0.5.

Using 10-fold cross validation, we obtain a test R2 of 85%. Installed PV base, socio demographic and incentives explain 85% of the variation of searches of "solarcity." This implies that "solarcity" is possibly acting as a proxy for these factors.

DISCUSSION AND FUTURE RESEARCH

Our analysis demonstrates that models uniquely based on Google searches of major installer in the area, such as SolarCity and Sunrun, can be effective in predicting PV adoption particularly when there's a large PV market such as California. As a matter of fact, pooled models based on both of these terms are able to explain 85% and 72% respectively of the variation in PV adoption in an independent dataset. The search-based models based on terms directly linked to PV such as "solar panel" did not perform as well as the name of installers.

While baseline models based on publicly available data outperformed search-based models, the search based models are still useful when baseline data is not available or hard to collect. Nonetheless, it will be worthwhile to further evaluate the robustness of search-based models by testing how well they can predict installations from the year 2017, which wasn't included in the analysis.

As for Connecticut, the google activities of searches related to PV is still very limited with only "solarcity" having a correlation coefficient with PV adoption of higher than 0.5. Still, the years pooled search-based model with the term "solarcity" had a weak performance. This could be explained by the fact that Connecticut has a small PV market, not enough for search terms to be predictive of adoption.

The search-based year fixed effect models performed very well for all terms and for both states. However, the fact that the search terms lose statistical significance when years fixed effect are added to the models implies that the strong performance of these models is associated to yearly factors rather than the search terms.

The only term that doesn't lose significance when adding years fixed effect is "solarcity" in California. This suggests that searches for "solarcity" are capturing variations that yearly factors are not capturing. This finding and the strong performance of the "solarcity" search-based model in the years pooled model in California imply that google searches of leading PV installers in large PV markets such as California can be a powerful predictive tool. In order to be confident in this conclusion, further research should test more than the one, California, large PV market tested in this paper.

The models combining both search terms and baseline data did not perform better than the baseline models. This suggests that searches of major PV installers do not add to the predictive power of baseline models, but rather might be capturing the same variation of baseline data, and thus can be proxying for incentives and socio-demographic data. Overall, the results of this study show that the potential realized by researchers in other domains can be realized by researchers interested in residential solar.

Our results inspire cautious optimism regarding the utility of search activity for predicting residential solar adoption. Future research must test/correct for the following limitations. First, our project was limited by the availability of Google Trends data at the state level. If more data becomes available, it will enable researchers to test an even greater volume of search terms at an even finer resolution. With enough state-level data, it would also be possible to test entirely new kinds of models or apply statistical learning techniques such that the model selects the best combination of terms. Second, SolarCity has merged with Tesla, so the search term "solarcity" may not be a good predictor of PV adoption (even in California) going forward.

Search activity data is only one of a seemingly infinite set of different kinds of online user data available to researchers and advertisers. Data acquired from social networking sites would provide a geographically targeted, multidimensional alternative to search data. It is our hope that our project will serve as a proof-of-concept and that future researchers will be able to specify the most correlated data streams, and construct uniquely predictive models. These future models may help researchers and firms better match supply and demand in the residential solar market.

REFERENCES

Della Penna, Nicolás, and Haifang Huang. 2010. “Constructing Consumer Sentiment Index for U.S. Using Google Searches.” Working Paper 2009–26. University of Alberta, Department of Economics. http://econpapers.repec.org/paper/risalbaec/2009_5f026.htm.

Drury, Easan et al. 2012. The transformation of southern California's residential photovoltaics market through third-party ownership. *Energy Policy*, Volume 42: 681–690.

Goel, Sharad, Jake M. Hofman, Sébastien Lahaie, David M. Pennock, and Duncan J. Watts. 2010. “Predicting Consumer Behavior with Web Search.” *Proceedings of the National Academy of Sciences*, 107 (41): 17486–90. doi:10.1073/pnas.1005962107.

Google Trends: trends.google.com. 2018

Mccallum, Malcolm L., and Gwendolyn W. Bury. 2013. “Google Search Patterns Suggest Declining Interest in the Environment.” *Biodiversity and Conservation*, 22 (6–7): 1355–67. doi:10.1007/s10531-013-0476-6.

Polgreen, Philip M., Yiling Chen, David M. Pennock, and Forrest D. Nelson. 2008. “Using Internet Searches for Influenza Surveillance.” *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America*, 47 (11): 1443–48. doi:10.1086/593098.

Rai, V., and Robinson, S. A. 2013. Effective Information Channels for Reducing Costs of Environmentally-friendly Technologies: Evidence from Residential PV markets. *Environmental Research Letters*, *8*(1), 014044.

Rai, V., Reeves, D. C., and Margolis, R. 2016. Overcoming Barriers and Uncertainties in the Adoption of Residential Solar PV. *Renewable Energy*, *89*, 498-505.

US Census Bureau. 2018. https://www.census.gov/programs-surveys/acs/

Vosen, Simeon, and Torsten Schmidt. 2011. “Forecasting Private Consumption: Survey-Based Indicators vs. Google Trends.” *Journal of Forecasting*, 30 (6): 565–78. doi:10.1002/for.1213.